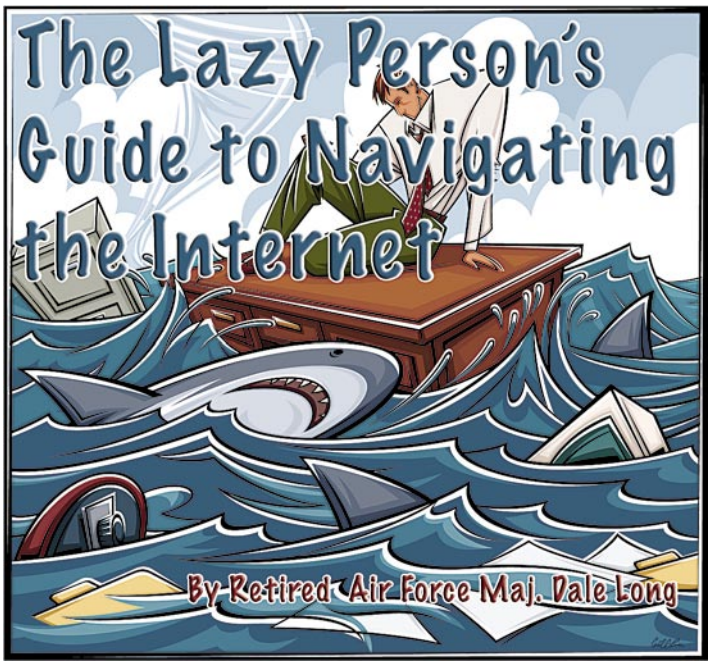

"No mercy, no power but its own controls it. Panting and snorting like a mad battle steed that has lost its rider; the masterless ocean overruns the globe."

– Herman Melville, from the novel "Moby Dick"

With more than 81 million host names in use as of May 2006, the Internet now resembles the vast ocean described by Herman Melville in 1851. In Melville's time, travel across an ocean was still an adventure. Now, through the near-magical quality of modern information technology, we have almost instantaneous, worldwide access to a vast ocean of information.

However, given the relative newness of this vast new cyber-ocean, many Internet surfers may find themselves far from shore without a compass. While all we have to do to return home is close our browser, it is still a bit disconcerting to find ourselves lost, adrift or even hijacked while doing something online.

So, in the interest of making the waters of the Internet less foggy and more navigable, in this issue we will look at how the Internet is organized, some of the navigation aids and other information sources available on the Net, and how to tell whether or not you can trust the site you are about to load.

Who's in Charge?

The Internet, along with the deepest ocean trenches and the outer reaches of the solar system, is one of the great modern frontiers for human exploration. As the Advanced Research Projects Agency Network (ARPANET) from 1969 until 1998, it was governed in various ways by the U.S. Department of Defense or associated contractors under the auspices of the Internet Assigned Numbers Authority (IANA) and other entities.

In 1998, perhaps in recognition of the Internet's transformation to a commercial entity, management of the Internet moved to a non-profit corporation sponsored by the Department of Commerce, the Internet Corp. for Assigned Names and Numbers.

ICANN is an internationally organized, non-profit corporation based in Marina del Rey, Calif. It is responsible for, among other things: managing Internet Protocol (IP) address space allocation; managing generic (gTLD) and country code (ccTLD) Top-Level Domain names; root server system management functions; preserving the operational stability of the Internet; and developing Internet management policy.

The most visible function of ICANN is its management of the Domain Name System (DNS). Every computer on the Internet has a unique IP address, a 32-bit number made up of four 8-bit "octets" that define every site on the Internet. For example, the IP address of the ICANN.org Web site is 192.0.34.163.

However, as most people have a hard time remembering arcane strings of digits, the DNS allows Web sites to use text as an alias for a numeric IP address, allowing us to type "www.icann.org" instead of the numeric IP address.

The principal value of the DNS is ensuring universal resolvability of Internet site addresses. This ensures that every Internet user, can access content from any site on the Internet. While there may be some governments that may not be entirely happy that their citizens can access allegedly unhealthy content via the Internet, ICANN and the Internet community have thus far successfully resisted having the Internet split up into segregated enclaves controlled by national or regional interests. The Internet remains an international resource, though with varying levels of monitoring, privacy and censorship depending on where you are.

What's in a Name?

Internet addresses are divided into groups of sites defined by domain names. "gTLD" is intended for use, at least in theory, by a particular class of organization. gTLDs were originally named for the types of organizations they represent, though some have become less restrictive over time. Let's start by looking at six gTLDs we are all probably familiar with.

.com – This domain is intended for commercial organizations, but anyone can apply for a dot-com address. There are more dot-com sites on the Internet than any other domain. The quality and reliability of these sites can vary widely, ranging from reputable sites associated with established companies to sites serving as fronts for phishing operators and online swindlers.

.edu – This domain is reserved for educational institutions. However, use of an dot-edu domain does not necessarily guarantee that the site belongs to an institution accredited by the U.S. Department of Education or equivalent foreign government agency.

.net – This domain was originally used to designate network infrastructures, but is now unrestricted. Commercial e-mail providers often use dot-net for their users' e-mail accounts *(e.g., Verizon.net, Adelphia.net, etc.)* possibly in an attempt to give the account more "net credibility" than a dot-com account.

.org – This domain was originally intended mainly for non-profit organizations that did not fit cleanly within the other gTLDs. However, like dot-net, the dot-org domain is now unrestricted.

.gov – This is a restricted domain reserved for the exclusive use of U.S. government agencies. ".mil" is similarly restricted for the exclusive use of U.S. military services and the Defense Department.

In addition to those six, the next set of sites you are likely to see are those assigned by country (ccTLDs), like ".ca" (Canada), ".ru" (Russia) or ".au" (Australia). Aside from these gTLDs and the ccTLDs, here are some lesser-known gTLDs:

.aero	for the air transport industry
.biz	for business use
.cat	for Catalan language/culture
.coop	for cooperatives
.eu	for the European community
.info	for informational sites, but unrestricted
.int	for international organizations established by treaty
.jobs	for employment-related sites
.mobi	for sites catering to mobile devices
.museum	for museums
.name	for families and individuals
.pro	for certain professions
.travel	for travel agents, airlines, hoteliers, tourism bureaus, etc.

The Internet is a big place, in a virtual sense. Netcraft.com, an Internet monitoring site, received responses from 81,565,877 sites in its May survey. According to Netcraft, the Internet grew by 909,000 sites from April to May and by 7.2 million hostnames from the beginning of the year through May. If you are keeping track that means the Internet gets a new hostname about every 3 seconds. Netcraft estimates the Internet will grow by 17 million hostnames this year.

Hostnames do not equal servers or pages. A site may have many servers and any number of pages. How many pages, you may ask? A site called the "WayBack Machine" (http://www.archive.org/web/web.php) has archived over 55 billion Web pages produced since 1996. To view them all you would have to view one page every second for the next 42,000 years. This begs the following question: How do we find anything in an ocean of information that mind-numbingly big?

Navigation Aids

Two types of sites help us navigate the Internet: portals and search engines. Portals are sites that collect and organize information and other functionality in your browser window based on preset conditions. The organization you work for probably has a portal of some type. Your Internet service provider (ISP) probably has a portal, and there are commercial Web sites like Yahoo.com, MSN.com and Google.com that anyone can use as a window to the Internet.

What distinguishes portals from other Internet sites is the amount of control you can exercise over what appears in your browser. My experience has been that commercial portals offer users a greater degree of customization than portals developed by companies or government agencies for their employees. I submit, however, that the popularity of a portal has a direct relationship to how much control users have over the content.

Humans like control. If I control my portal space, I am not going to clutter it with advertisements or press releases. I'm going to include stuff I am actually interested in and use. I will accept some content from the portal owner, but if I cannot control the majority of my home page space, I will go elsewhere.

The commercial portal that is currently my home page on every computer I use allows me to create multiple pages with news feeds, links to government, financial and technology sites, Web comics, and search sites. It is a window that satisfies my personal and professional needs. The trade-off is that the portal manager can show ads in the top banner and in a side column.

Search Me

Portals organize things based on preset conditions. When we need to find something new, we use a search engine. The first generation of Internet search tools started with "Archie," *(the word "Archive" without the letter "v")* created in 1990 by Alan Emtage, a student at McGill University in Montreal. However, Archie did not search though file content. It just downloaded the directory listings of all the files located on public anonymous File Transfer Protocol (FTP) sites and created a searchable database of filenames.

In 1991, students at the University of Minnesota developed "Gopher" *(named after the school's mascot)* which indexes plain text documents. Gopher is a distributed document search and retrieval network protocol designed for the Internet. Its purpose was similar to that of the World Wide Web. The Web has almost completely displaced Gopher. However, there are still a few active Gopher sites in existence, including one at the Smithsonian Institution.

Two other programs, apparently developed by people who missed the memo that Archie wasn't named after a comic strip character, were "Veronica" and "Jughead," which searched the files stored in Gopher index systems. Veronica *(Very Easy Rodent-Oriented Net-wide Index to Computerized Archives)* provided a keyword search of Gopher menu titles. Jughead *(Jonzy's Universal Gopher Hierarchy Excavation and Display)* obtained menu information from Gopher servers.

Then the World Wide Web tsunami swept over the Internet, changing it forever. The proof of concept for Web searching debuted in 1993 with Aliweb *(Archie Like Indexing for the Web)*. The first well-known full-text search engine on the Web was WebCrawler in 1994, soon joined by Infoseek and Lycos.

AltaVista and Excite appeared in 1995, with Dogpile, Inktomi and Ask.com rounding out the second generation of Internet search engines in 1996.

These full-text search engines held their own for a while, but eventually fell victim to three things: the Web started getting

too big for their technology; the dot-com bubble burst; and someone built a better search mousetrap.

In 1998, the beta version of Google appeared on the Web. While Google also uses text indexing, it pioneered two features that gave it an edge over other browsers: link popularity and PageRank. Link popularity measures the quantity and quality of Web sites that link to pages with content that meets your search criteria. While text indexing can measure how a page meets search criteria quantitatively, link popularity is a qualitative measure of "off-the-page" criteria.

The theory is if a page is important or useful, other sites will have links to it, and pages with little or no value will have fewer citations. Link popularity analyzes how many other sites link to the target page and cross-references that with the linking site's reputation. It is the Web equivalent of "word-of-mouth" referrals.

PageRank is the heart of Google. According to Google, *"PageRank relies on the uniquely democratic nature of the Web by using its vast link structure as an indicator of an individual page's value. In essence, Google interprets a link from page A to page B as a vote, by page A, for page B. But, Google looks at more than the sheer volume of votes or links a page receives. It also analyzes the page that casts the vote. Votes cast by pages that are themselves 'important' weigh more heavily and help to make other pages 'important.'"*

The combination of these two features allowed Google to generate more accurate search results than any other search engine on the Web at the time. Other search engines have attempted to copy its methods, but Google still has approximately 80 percent of the search engine market through user trust in their results.

However, even Google does not claim to index the entire World Wide Web — just around 20 billion pages. That leaves room for specialized search sites based on a concept known as vertical search. Google is a horizontal search engine; it attempts to index across as much of the Web as possible. Vertical search engines specialize in content areas, like travel, real estate or retail sales, and only include sites that match their special interest criteria.

As the Web grows larger, it is likely it will grow beyond the capability of any single horizontal search engine to keep up. What we may have in another 10 years are vertical search engines that work in particular content areas or domains and meta-search engines that send our queries out to multiple vertical and horizontal search engines and aggregate the results. For example, WebCrawler is now a meta-search engine.

Trust, but Verify

This brings us to a few closing thoughts on the value, authenticity, and reliability of what is displayed in our portals or search engines. How can you tell if a Web site is both legitimate and useful?

The first indicator is the domain name. If you are visiting a dot-gov or dot-mil site, it is a pretty safe bet that the content is legitimate. With any other domain, however, you take your chances. I am more inclined to trust dot-edu, dot-org or dot-net domains than dot-com or dot-ru, though I do look for independent verification.

Here's a quick quiz. Which of the following links are what they appear to be?

1. http://travelocity.com/
2. http://paypal-email.com/login.htm/
3. http://www2.usairways.com/
4. http://www.ebay.com@64.236.24.12
5. http://www.email.citicards.com/

Now check your answers. How did you do?

No. 1 is a legitimate link to Travelocity.

No. 2 was once used as a phishing link to a fake PayPal site that would capture your account login and give the phishers access to your account. It is no longer active.

No. 3 is a legitimate US Airways link.

No. 4 is a phisher-style address that attempts to redirect you to a different site. In this example, the numeric IP address after the @ will attempt to redirect you to a site that doesn't require authentication, for example, CNN.com. If the destination site is a phishing site built to require authentication and accept "www.ebay.com" as valid data, you would get no warning about the redirect, and you could be looking at something that looks like eBay — but isn't. This site is now blocked on many networks.

No. 5 is a trick question. Yes, this is a legitimate CitiBank site address. But clicking on this link in a recent CitiBank e-mail actually took you to a different address. Disguising links with a different address label is a common phishing trick, both in e-mail and on Web sites. Most programs with the ability to activate Web links will at least briefly display the actual link address when your mouse cursor pauses over a link.

I highly recommend making sure where any link is actually going. Finally, you should report rogue links to your ISP for everyone's protection.

Aside from phishing and other technical tomfoolery, there is another trust issue: *Is the content on any given Web site useful or truthful?* Unfortunately, there is no way to check this with technology. As with any source of information, like newspapers, television news or talk radio, we still have to use good judgment on the content.

Use this old adage as a good rule of thumb: *"Believe half of what you see and none of what you hear."* Of course, we still have to decide which half, but at least we have a 50 percent chance.

Until next time, Happy Networking!

Long is a retired Air Force communications officer who has written regularly for CHIPS since 1993. He holds a Master of Science degree in Information Resource Management from the Air Force Institute of Technology. He is currently serving as a telecommunications manager in the U.S. Department of Homeland Security. CHIPS